Is Email Business Dying?: A Study on Evolution of Email Spam Over Fifteen Years★

De Wang[1,†], Danesh Irani[1,*], and Calton Pu[1,†]

[1]College of Computing, Georgia Institute of Technology, Atlanta, Georgia 30332-0765

Abstract

With the increasing dedication and sophistication of spammers, email spam is a persistent problem even today. Popular social network sites such as Facebook, Twitter, and Google+ are not exempt from email spam as they all interface with email systems. While some report predicts that email spam business is dying due to the decreasing volume of email spam. Whether email spam business is really dying is an interesting question. In this paper, we analyze email spam trends on Spam Archive dataset, which contains 5.5 million spam emails over 15 years (1998 – 2013). We statistically analyze emails contents including header information (e.g. content type) and embedded items (e.g. URL links). Also, we investigate topic drift using topic modeling technique. Moreover, we perform network analysis on sender-to-receiver IP routing networks. Our study shows the dynamic nature of email spam over one and a half decades and demonstrate that the email spam business is not dying but more capricious.

Keywords: email, spam, evolution

Received on 21 January 2014; accepted on 11 April 2014; published on 27 May 2014

doi:10.4108/cc.1.1.e3

1. Introduction

As a method to communicate both for individuals and businesses everyday, email is also used as an information management tool [1]. What started primarily as a person-to-person communication medium has spread widely to one-to-many (e.g. mailing-lists) and many-to-one (e.g. forwarded traffic) communication medium [2]. As social media has grown dramatically, email also enhances the functionality provided by them. For instance, users are sometimes given pseudo-email addresses which can be used to receive emails on the social networks as well as email can sometimes be used to interact with the social networks using specially crafted email addresses.

Due to the convenience and popularity of email system, malicious users also take it as a major target to launch Denial of Information (DoI) attacks [3]. Spam pollution is one kind of DoI attacks, which prevents users from finding non-spam content. Spam is unsolicited and unrelated content sent to users, which most commonly is associated with email, but also applies to several different domains including instant messaging, websites, and Internet Telephony [4–8]. Spam degrades a user's experience as, by definition, it is an annoyance and gets in the way of users consuming non-spam content.

In August 1998, Cranor et al. [9] described the rapidly growing onslaught of unwanted email and since then the volume of spam has grown even more as the amount of all email sent has grown exponentially. Constituting an annoyance, email spam has increased to as much as 90% today [10] from approximately 10% of overall mail volume in 1998, which results in an enormous burden on the thousands of email service providers (ESPs) and millions of end users on the Internet [11].

In addition to being on the receiving side of spam, ESPs need to invest in developing filters to combat the spammers and likewise spammers evolve to avoid spam filters. The co-evolution nature of spammers and spam filters is an "arms-race", which has resulted in numerous publications employing

★Extended version of conference paper at CollaborateCom 2013.

*The work done on the paper was during Danesh Irani's PhD at Georgia Institute of Technology. He is software engineer at Google now.

†Corresponding authors. Email: wang6@gatech.edu, calton.pu@cc.gatech.edu

adversarial strategies to tackle the spam problem [12–14]. Pu et al. [15] and Fawcett [16] developed techniques for characterization and measurement of email spam trends and researchers have also examined other types of spam including phishing [17] and Web spam [18]. In addition, Guerra et al. [19] compared the effectiveness of old and recent filters over old and recent spam to obtain spam trends on email spam dataset.

In this paper, we investigate the trends of email spam in terms of content, topics, and sender-receiver network over 15 years by performing an evolutionary study on the Spam Archive dataset [20]. We aim to answer the question of whether the email spam business is dying (also, as identified by our title). More concretely, we make the following contributions:

- We perform a long-term evolutionary study on a large email spam dataset, which includes statistical analysis, topic modeling and network analysis.
- We demonstrate the changes of email spam over time with respect to contents and spammer behaviors.
- We prove that email spam business is not dying but is becoming sophisticated by the evolutionary study on large scale real data.

The remainder of the paper is organized as follows. We motivate the problem further in Section 2. Section 4 introduces the Spam Archive dataset used in our study. Section 5 presents the analysis performed on the dataset and findings derived from the results. Section 6 discusses the future of email spam business and the limitations of our study. We talk about related work in Section 3 and conclude the paper in Section 7.

2. Motivation

The paper is inspired by an article by Kaspersky labs [21] named “The dying business of email spam” [22], which stated that “Spam email is on the wane. And no one on God’s green Earth is going to miss it”. The conclusions were based on their annual report [23] citing that the share of spam in email traffic decreased steadily throughout 2012 to hit a five year low.

We are excited by the decline in the volume of email spam but it also raises the question as to whether the email spam business is dying and will continue to decline. Besides the volume change, we also consider the quality of email spam and the impact, which may be constituting a new trend of email spam business. For instance, spammers may post email spam in a more complicated way using spoofed email addresses and changing email relay servers. Those kind of email spam may slip away under the inspection of spam filters. Thus, it motivated us to investigate the evolution of email spam using advanced techniques such as topic modeling and network analysis. We try to find out the real trend of email spam business through email content, meta information such as headers, and sender-to-receiver network over a long period of time.

3. Related Work

3.1. Email Spam Detection

Email spam detection has been studied by lots of researchers in different directions. For instance, Carreras et al. [24] applied boosting trees to filter out email spam. Wang et al. [25] used heuristic feature selection techniques to improve the performance of email spam filtering. Chan et al. [26] co-trained with a single natural feature set in email classification. Liu et al. [27] adopted multi-field learning for email spam classification. Sculley et al. [28] used relaxed online SVMs for email spam filtering. Besides those machine learning techniques, more researchers tried other kinds of detection methods. Attenberg et al. [29] introduced collaborative email spam filtering with the hashing trick. Balakumar et al. [30] offered ontology based classification of email. Dasgupta et al. [31] combined similarity graphs to enhance email spam filtering. Jung et al. [32] used DNS black lists and spam traffic to detect email spam. Ramachandran et al. [33] filtered email spam with behavioral blacklisting. Clayton et al. applied extrusion detection in stopping email spam by observing distinctive email traffic patterns. Xie et al. [34] provided an effective defense approach against email spam laundering. Additionally, researchers also have used email spam to help detecting other types of spam. For instance, Zhuang et al. [35] developed an approach to map botnet membership using traces of spam email. Webb et al. [36] identified an interesting link between email spam and Web spam and used it to extract large Web spam samples from the Web. Wang et al. [37] demonstrated the relationship among different formats of social spam including user profile spam, message spam and Web spam, in which message spam contain email spam.

3.2. Information Retrieval on Email Data

Another focus of researchers is information retrieval on email data. Bird et al. [38] constructed social networks of email correspondents to address some interesting questions such as the social status of different types of participants and the relationship of email activity and other activities. McCallum et al. [39] illustrated experimental study on Enron and academic email to discover topic and role in social networks from emails, in which the model builds on Latent Dirichlet Allocation (LDA) and the Author-Topic (AT) model.

Culotta et al. [40] presented an end-to-end system that extracts a user's social network and its members' contact information given the user's email inbox.

3.3. Evolutionary Study of Spam

Research work on evolutionary study of spam is close to this paper [41, 42]. Pu et al. [15] presented a study on dataset collected from Spam Archive and focused on two evolutionary trends: extinction and existence. Irani et al. [17] studied the evolution of phishing email messages and classified them into two groups: flash attacks and non-flash attacks. Wang et al. [18, 43] compared two large Web spam corpus: Webb spam corpus 2006 and Webb spam corpus 2011 and shown the trending of Web spam. Chung et al. [44] and Fetterly et al. [45] also have done intensive study on evolution of web spam. Guerra et al. [19] investigated how the popularity of spam construction techniques changes when filters start to detect them and determined automatically techniques that seemed more resistant than others. The evolution of spamming techniques shows the increasing sophistication of spammers. Our work focuses on tactics changes of email spam over time and inspires more researchers to work on email spam detection collaboratively.

4. Data Collection

In this section, we introduce the Spam Archive dataset and show the overview of the dataset used in our study.

Spam Archive dataset [20] is collected by Bruce Guenter since early 1998 using honey-pot addresses. The project is still ongoing with monthly releases of new email spam. Since it provides a continuous long-term email spam data source from a consistent source, it is an excellent dataset for our investigation into spam trends. The volume of email messages received over the 15 years is shown in Fig. 1, with the date on the x-axis and log-scale volume of email messages received per month on the y-axis. From the figure we see that email spam volume grows steadily over time. For the spike of email spam during 2006, Bruce Guenter has attributed this to one of the spam traps having a wild-card address which received increasingly large amounts of spam which was subsequently disabled after 2006, since most of the spam was duplicates of other spam received.

Besides showing the trend of overall volume of email spam, we also present the volume changes monthly for different years in Fig. 2, with the month of the year on the x-axis and the log-scale volume of spam messages per month on the y-axis. It shows volume trends over the previous 15 years. The volume of email spam is not always increasing over time such as the email spam volume changes during 1999. Some years' volumes also shows fluctuations over time. For instance, during 2002, the volume first went up in May and decreased dramatically afterward until July. Several factors may have contributed to this change such as new strategies used by spammers (e.g. image spam is introduced in emails), improved spam filters (e.g. URL analysis tool is adopted) and even political influence from governments (e.g. Electronic Communications and Transactions Act, 2002 [46]). We investigate the details and potential reasons of these changes in more detail in the following sections.

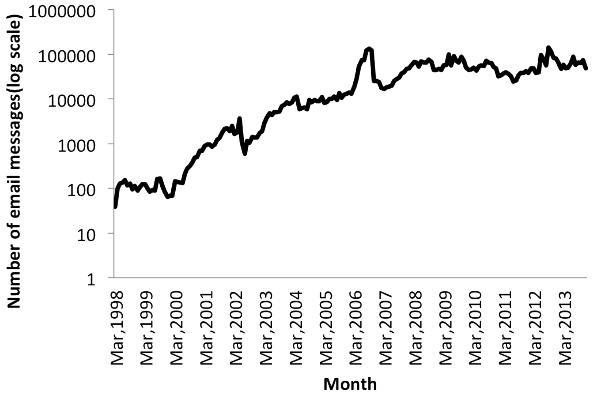

Figure 1. Number of email messages (per month) over time

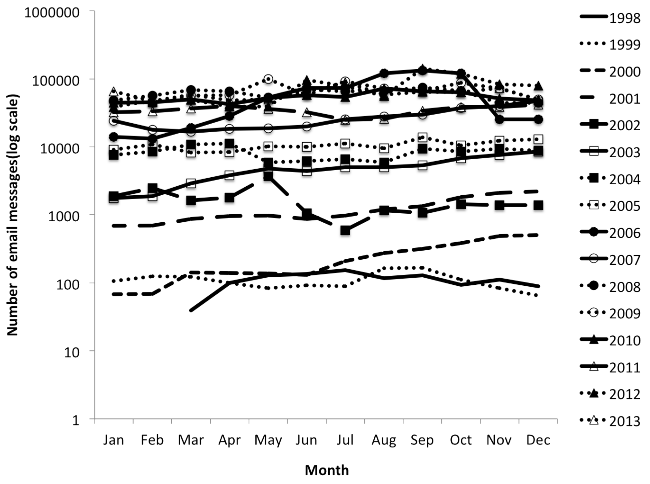

Figure 2. Number of email messages in month order for different years

5. Data Analysis

In this section, we start with content analysis of Spam Archive dataset, followed by topic modeling and network analysis.

5.1. Content Analysis

The two main types of email message content are “Text" and “Multipart". Messages in type “Text" are simple text messages while messages in type “Multipart" have parts arranged in a tree structure where the leaf nodes are any non-multipart content type and the non-leaf nodes are any of a variety of multipart types [47]. To have a better sense of the distribution of main types in email spam, we show the main type distribution in different years in Fig. 3.

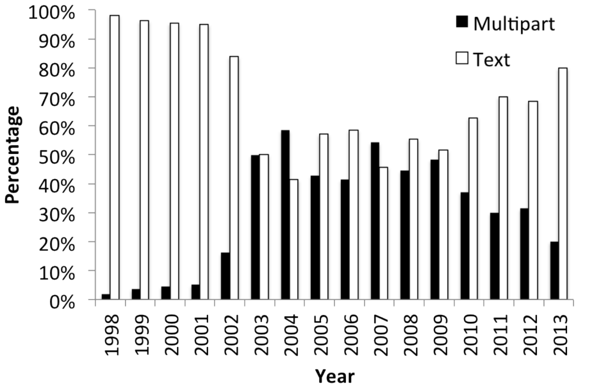

Figure 3. The distribution of main types of message content

Fig. 3 demonstrates that the distribution of two main types in our dataset changed over time. For instance, before 2003, more email spam had the message format in the main type “Text". After that, the two main types almost occupied the same percentage until 2010. The new trend is that email spam is using more messages in main type “Text" (e.g. the percentage of email spam in main type “Text" is about 80% for the year of 2013).

Next thing we are interested in is the embedded items in email spam such as HTML web page, images, and URL links. After scanning all email spam in our dataset, we present the distribution of embedded items in email spam over time in Fig. 4.

Fig. 4 shows that low percentage of email spam, which was always less than 5% in our dataset, contained image attachments. On the contrary, more email spam had embedded HTML web pages and URL links. But the percentages of email spam containing HTML web pages and URL links changed dramatically over time. Several peaks and valleys appeared over 15 years in the Fig. 4. For instance, HTML pages had peaks in 2003, 2007, and 2009 and valleys in 2006 and 2008. While for URL links, peaks appeared in 2004, 2008 and 2012 and valleys appeared in 2006 and 2011. Since HTML page normally carries URL links, they should have similar fluctuations along the time. However, we observe that an exception occurred after 2011. The percentage of email spam containing HTML web pages decreased suddenly after 2009. While the percentage of email spam containing URL links dropped down along with HTML web pages until 2011 and it increased sharply afterwards. One possible reason is that more URL camouflage techniques, which are quite efficient in avoiding spam filters, appeared such as shortened URLs and hidden URLs in recent years. To investigate further the trend of URL links, we aggregate all URL links on a yearly basis for email spam that contain URL links and show the cumulative distribution of URL links in email spam in Fig. 5 (1998 – 2013).

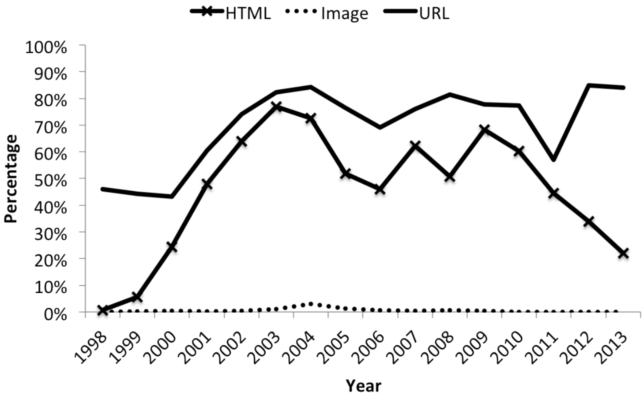

Figure 4. The distribution of embedded items in email spam over time

Fig. 5 shows the number of URL links for the majority of email spam is below 10. Only a small portion of email spam have more than 1,000 URL links which may be embedded in different depths of email messages. Even though the densities of URL links in email spam changed variously, email spam contained more and more URL links over time.

Through the analysis, we obtain the following observation (**Observation I**):

- In terms of percentage, very few image embedded items appear in the email spam. One possible reason is that email system, such as the Gmail system, adopts new policy to automatically hide the images in emails unless user chooses to display them.
- Email spam contains more text and more URL links in recent years. Many URL links are legitimate URL links such as Facebook or Google official website. Spammers use legitimate URL attack to avoid detection and increase the cost of filtering at least since the spam filter needs to go through all the URL links in email to distinguish the message from legitimate ones.

In addition to looking into embedded items, we also investigate the top n-grams in email spam over time. The tool we used for obtaining n-grams of email spam is

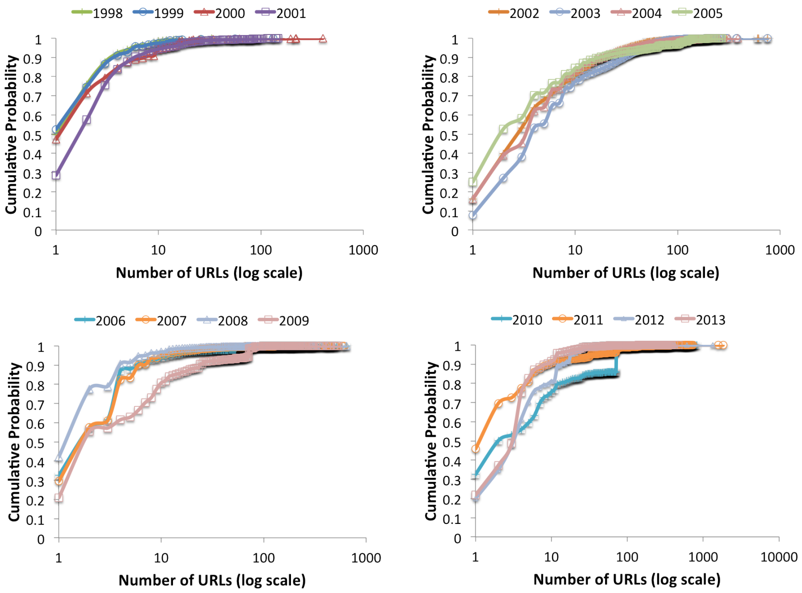

Figure 5. Cumulative distribution of URL links in different years

Perl's module Text::Ngrams [48]. First, we need to clean our dataset by filtering out stop words and striping out HTML tags. And then we calculate top-10 n-grams (n ranges from 1 to 3) on a monthly basis over 15 years. Due to space limit, we only list the top-10 n-grams starting from June 1998 to June 2013, which is shown in Table 1.

In Table 1, $\langle N \rangle$ denotes any number sequence. Top-10 n-grams set contained different words or word sequences along the time, showing different topics as well. For instance, the n-grams set in June 1998 tells us that the email spam was advertising fake dental services using attractive words such as "free", "nationwide near", and "month save average". The n-grams set in June 2003 was about marketing and market leaders leading people to click external URL links. The n-grams set in June 2008 was about DASS (Defensive Aids Sub System) [49] which is a fighter system from European countries. After checking the original email, it is a trap news or game to attract the email receivers to enter into. The n-grams set in June 2013 was more related to new media announcement and membership registration. Thus, we have the observation (**Observation II**):

- The content of N-gram sets changed over time. Spammers try to obtain users' interests by keeping the content up-to-date and attractive. Also, frequent changes of contents make email spam hard to be detected by spam filters based on content analysis.

Moreover, the differences indicate the topic drift in email spam over time (e.g. from fake advertising to fake registration services). To learn more about the topic drift of email spam, we will apply topic modeling on the dataset next.

5.2. Topic Modeling

Topic modeling is defined as a technique that looks for patterns in the use of words and it is an attempt to inject semantic meaning into vocabulary, in which a "topic" consists of a cluster of words that frequently occur together [50]. The tool we used in our topic modeling is a machine learning toolkit for language named "MALLET" [50]. It provides an efficient way to build up topic models based on Latent Dirichlet Allocation model (LDA) [51].

To simplify the illustration, we set up the number of topics to 10 in the data processing. After the calculation, we obtain the word (also called term) lists associated with topics and topic composition for different months over time, which is shown in Table 2 and Fig. 6.

In Table 2, it shows the topic name and the samples of the most related terms. After the topic modeling, we only have the word or term clusters for each topic which has not been labeled. Based on associated terms with each topic and experience with email spam, we label the topics as "Account Information", "Order Information", "Business News", "Sales News", "Adult Product", "Software Product", "Official News", "Free Product", "Medical Product", and "Newsletter" separately. Due to the space limit, we just list sample of most related terms for each topic in Table 2.

Fig. 6 shows the topic drift in our dataset. We observe that the popular topics drifted along the time. Before 2004, the topic "Business News" was the most popular topic in email spam. After that, the most popular

Table 1. List of top-10 n-grams every 5 years on a monthly basis (n ranges from 1 to 3)

June, 1998	June, 2003	June, 2008	June, 2013
dental	$\langle N \rangle$	$\langle N \rangle$	$\langle N \rangle$
free	click	euro	important
plan	email	dass	garden
$\langle N \rangle$	information	online	class
details	bait	http	email
call	mail	mail	media
please	free	original	screen
doctor	message	super	dark
dentistry	work	time	right
procedures	please	active	registration
plan free	$\langle N \rangle \langle N \rangle$	$\langle N \rangle \langle N \rangle$	$\langle N \rangle \langle N \rangle$
teeth whitening	email bait	euro euro	garden $\langle N \rangle$
nationwide near	august $\langle N \rangle$	super active	$\langle N \rangle$ garden
waiting periods	market information	active euro	media screen
root canals	world leader	tabs doses	important media
details june	auction records	kinder dass	important important
dental procedures	remove email	autopilot dass	dark skin
canals crowns	reply message	original stress	screen class
doctor locator	link work	stress angst	class important
polishing fillings	leader market	angst dass	rights reserved
sealants prevent cavities	$\langle N \rangle \langle N \rangle \langle N \rangle$	$\langle N \rangle \langle N \rangle \langle N \rangle$	garden $\langle N \rangle$ garden
doctor locator number	world leader market	euro euro euro	$\langle N \rangle$ garden $\langle N \rangle$
crowns dentures braces	leader market information	active euro euro	$\langle N \rangle \langle N \rangle \langle N \rangle$
problems qualify waiting	case link work	super active euro	important media screen
month save average	demander plus figurer	dass kinder dass	media screen class
receive optical plan	allow mail removed	dass autopilot dass	important important media
call $\langle N \rangle$ please	removed thank operation	dass dass kinder	class important media
canals crowns dentures	modifier sera effective	autopilot dass dass	screen class important
optical plan free	message modifier sera	original stress angst	limited become member
plan receive optical	effective coop demander	kinder dass super	become member soon

Table 2. List of topics and associated terms

Topic Name	Samples of Most Related Terms
Account Information	email important pass check account address information
Order Information	click message privacy online policy information address view order receive required
Business News	click information price free professional time link business work
Sales News	price life money make time today offer year online real world women retail deal credit
Adult Product	world price penis back people product degree patch life make great experience enlarge
Software Product	price professional click software company copy softwares read suite online site office
Official News	united states world state national city government international people
Free Product	online pills price click quality save products email item prices service offer free
Medical Product	generic save price time products medications order pharmacy home service product
Newsletter	mail click email privacy newsletter message receive view offers link subscribed

topic changed more frequently than before. First, the most popular topic changed to "Software Product" for around a year. And then it changed back to the topic "Business News" again. And later on, the most popular topic changes happened in the following order: "Adult Product", "Free Product", "Sales News", "Free Product", "Newsletter", "Official News", "Order Information", "Medical Product", and "Account Information". For each topic, it contains certain features that are attractive to certain group of users. For instance, topic "Free Product" is more attractive to users who like free stuff. Topic "Medical Product" is more attractive to users

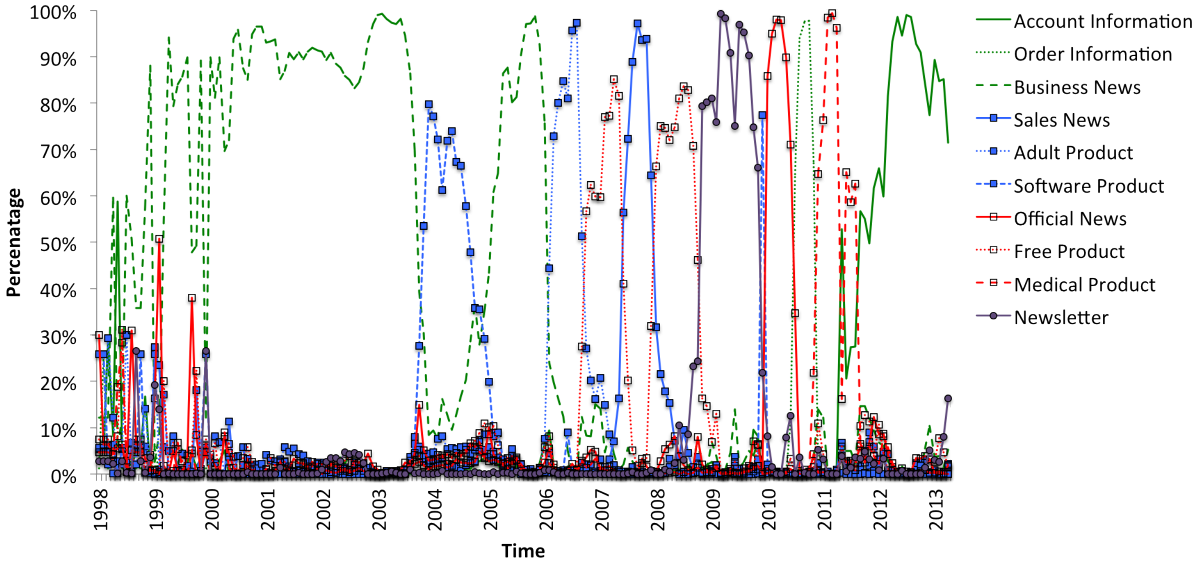

Figure 6. Topic drift in time order (time unit: month)

who need medical service or special medical products. Topic "Sales News" and "Order Information" are more attractive to users who like shopping. Meanwhile, as social media have interfaces with email systems normally and gain increasing popularity, email spam which have the content related to social media are growing rapidly. For instance, by investigating the content of email messages which belonging to the recent most popular topic "Account Information", we observe that a lot of email spam have associations with social media. One example is that social media account registration email spam which contains spam URLs that camouflaged as confirmation URL links. Another example is social media account notifications. For example, it informs you that your account has been changed by someone and needs immediate action to reset the password, followed by the spam URL links. Thus, one possible reason why the topic "Account information" becomes popular is that a lot of spammers try to impersonate the support team of social media to steal sensitive information, such as credential and credit information, or lead users to spam or phishing web pages for further actions. Thus, we conclude our observation as follows (**Observation III**):

- The topic "Business News" dominated in earlier years while the topic "Account Information" dominates recently. Topic drift happened frequently between 2004 and 2011. Meanwhile, lots of social engineering attacks are launched in later email spam.

5.3. Network Analysis

Besides content analysis and topic modeling, we also try to find out the sending behavior changes of spammers over time through analyzing the routing network between sender and receiver. Before entering into the detail of network analysis, we will talk about data processing and some findings during the process.

For the data processing, we need to process the headers of email message to obtain the information about routing between sender and receiver. The headers which are related to the routing info are "From", "To", "CC", "BCC" and "Received". The header "From" and "To" provide the sender and receiver email addresses. The header "CC" and "BCC" show the recipient lists in carbon copy and blind carbon copy mode. The header "Received" contains routing information from sender and receiver. First, we look into the headers "From" and "To" and intend to use them to extract the sender-to-receiver network. However, the fact is that we cannot use them in our study since most of the messages in the dataset contain forged "From" headers in one form or another, which is also mentioned in the Spam Archive dataset homepage. Although "From" header should not be trusted, we still extract top-10 domains from the "From" header to find out what are those popular domains used by spammers to set up social engineering traps for users. It is hard for users to recognize fake senders based on senders' email address especially when the email address is belonging to the domains they trust. The list of top-10 domains is shown in Table 3.

Table 3. List of top-10 domains

1998	1999	2000	2001
hotmail.com	yahoo.com	yahoo.com	hotmail.com
yahoo.com	hotmail.com	hotmail.com	yahoo.com
msn.com	aol.com	earthlink.net	excite.com
usa.net	usa.net	aol.com	msn.com
earthlink.net	ibm.net	usa.net	aol.com
att.net	msn.com	excite.com	btamail.net.cn
aol.com	iname.com	mail.com	earthlink.net
mailexcite.com	hotbot.com	bigfoot.com	mail.com
juno.com	bigfoot.com	email.com	pacbell.net
prodigy.com	mailcity.com	postmark.net	mail.ru
2002	**2003**	**2004**	**2005**
yahoo.com	yahoo.com	yahoo.com	yahoo.com
hotmail.com	hotmail.com	hotmail.com	hotmail.com
aol.com	aol.com	msn.com	msn.com
msn.com	msn.com	yahoo.co.kr	yahoo.co.kr
excite.com	artauction.net	aol.com	gmail.com
link2buy.com	earthlink.net	attbi.com	yahoo.co.jp
eudoramail.com	excite.com	yahoo.co.jp	163.com
flashmail.com	artaddiction.com	excite.com	msa.hinet.net
netscape.net	juno.com	seznam.cz	mail.com
btamail.net.cn	artists-server.com	netscape.net	126.com
2006	**2007**	**2008**	**2009**
yahoo.co.jp	yahoo.com	dyndns.org	dyndns.org
hotmail.com	dyndns.org	yahoo.com	homeip.net
mail.ru	hotmail.com	adelphia.com	untroubled.org
0451.com	yahoo.co.jp	hotmail.com	gmail.com
em.ca	paran.com	gmail.com	hotmail.com
yahoo.com	gmail.com	wikipedia.org	yahoo.com
0733.com	163.com	earthlink.net	untroubled.org
aol.com	msn.com	att.net	ezmlm.org
infoseek.jp	msa.hinet.net	163.com	em.ca
msn.com	so-net.ne.jp	cox.net	mail.ru
2010	**2011**	**2012**	**2013**
dyndns.org	yahoo.com	yahoo.com	yahoo.co.jp
yahoo.com	dyndns.org	garden.md	li-brooz.jp
homeip.net	ymail.com	yahoo.co.jp	yahoo.com
untroubled.org	gmail.com	ageha.cc	mixi1mega.biz
untroubled.org	mail.ru	peach.6060.jp	netstar-inc.co.uk
ezmlm.org	msn.com	ts5558.com	garden.md
em.ca	bk.ru	momoiro.cc	for-dear-2013.mobi
comcast.net	qip.ru	koikoilkoii.com	wakuwaku06.info
gmail.com	list.ru	wakuwaku-happy.net	greemmix.info
pfizer.com	aol.com	get-c.com	docomo.ne.jp

From Table 3, we observe that several popular email domains are used by spammers such as “yahoo.com", “hotmail.com", “msn.com", and “gmail.com". Also some top domains are related to receiver domains such as “untroubled.org" and “dyndns.org". It reveals that spammers were camouflaging themselves coming from the same domains as the users’ domains. In addition, some domains in the top-10 list are from countries outside US such as “163.com" which is the largest email service domain in China. In 2013, the top domains list contains more special domains such as “.biz" which is intended for registration of domains to be used by businesses and “.mobi" which is used by mobile devices for accessing Internet resources via the Mobile Web. It indicates that spammers were spoofing the sender addresses targeting business and mobile users.

Meanwhile, it proves that spammers recognize the trend of information flow in the Internet and evolve to take advantage of the trending.

Next, we investigate the header “CC" and “BCC" in email message to know whether spammers use those functions to spread email spam. The trends of “CC" and “BCC" are shown in Fig. 7.

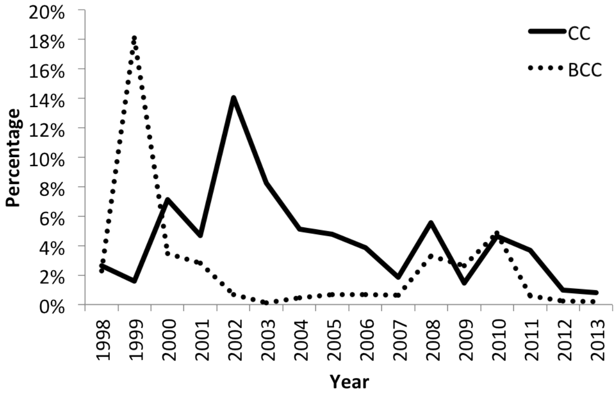

Figure 7. Cc and Bcc trends

Fig. 7 shows that spammers used more “CC" and “BCC" in the early years (1999-2004) and less in the recent years (2011-2013). One possible reason is that most spam filters have taken the number of “CC" and “BCC" as important features to detect spam [52]. Meanwhile, people become alert to email message which contains a long recipient list in the header “CC" and “BCC" so that this type of email spam lost markets gradually.

Thus, we conclude the observation as follows (**Observation IV**):

- Fields “FROM" and “TO" cannot be trusted. Meanwhile, they are used in social engineering attacks to camouflage email spam as emails from legitimate domains.
- Spammers use less CC and BCC now. Besides, they are also easy to be forged. So, they cannot be used in our network analysis.

Based on observations above, we realize that the header “From", “To", “CC", and “BCC" are not helpful in extracting routing network from email spam. To have a better understanding of the changes in terms of spammers' behaviors, we still need to find a way to extract the real sender and the routing information.

The header “RECEIVED" provides us the routing information such as hops' IP addresses between sender and receiver. Here is one example “RECEIVED" field in email header shown in Figure 8:

```
Return-path: <sender@senderdomain.tld>
Delivery-date: Wed, 13 Apr 2011 00:31:13
+0200
(3) Received: from mailexchanger.recipient
domain.tld([ccc.ccc.ccc.ccc]) by mailserver.
recipientdomain.tld running ExIM with
esmtp id xxxxxx-xxxxxx-xxx; Wed, 13
Apr 2011 01:39:23 +0200
(2) Received: from mailserver.senderdomain.
tld ([bbb.bbb.bbb.bbb] by mailexchanger.
recipientdomain.tld with esmtp id xxxxxx-
xxxxxx-xx for recipient@recipientdomain.tld;
Wed, 13 Apr 2011 01:39:23 +0200
(1) Received: from senderhostname [aaa.
aaa.aaa.aaa]} by mailserver.senderdomain.tld
 with esmtpa (Exim x.xx) (envelope-from
<sender@senderdomain.tld) id xxxxx-
xxxxxx-xxxx for recipient@recipientdomain.tld;
Tue, 12 Apr 2011 20:36:08 -0100
Message-ID: <xxxxxxxx.xxxxxxxx@senderdomain.
tld>
Date: Tue, 12 Apr 2011 20:36:01 -0100
X-Mailer: Mail Client
From: Sender Name <sender@senderdomain.tld>
To: Recipient Name <recipient@recipientdomain.tld>
Subject: Message Subject
```

Figure 8. Example of “RECEIVED" field in email header

Due to that the “RECEIVED" field is hard to be forged, we will use it to extract sender-to-receiver IP routing information and construct routing network. The tool we used in extraction is the email module in Python [53] and the network analysis tool is the open source network visualization software Gephi [54].

During the process of extracting networks, we also collect two extra features: average hops between sender and receiver and the Geolocation distribution of sender IP addresses. The list of average hops and the Geolocation distribution of sender IP addresses over time are shown in Fig. 9 and Fig. 10 respectively.

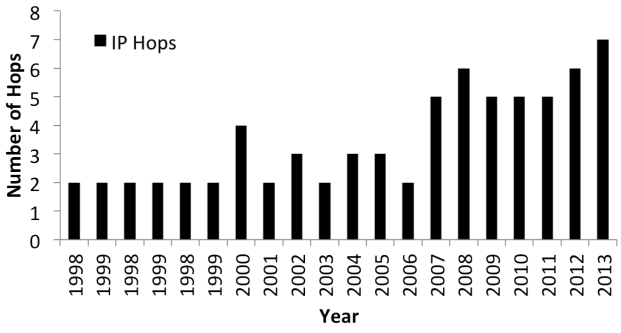

Figure 9. Average hops between sender and receiver

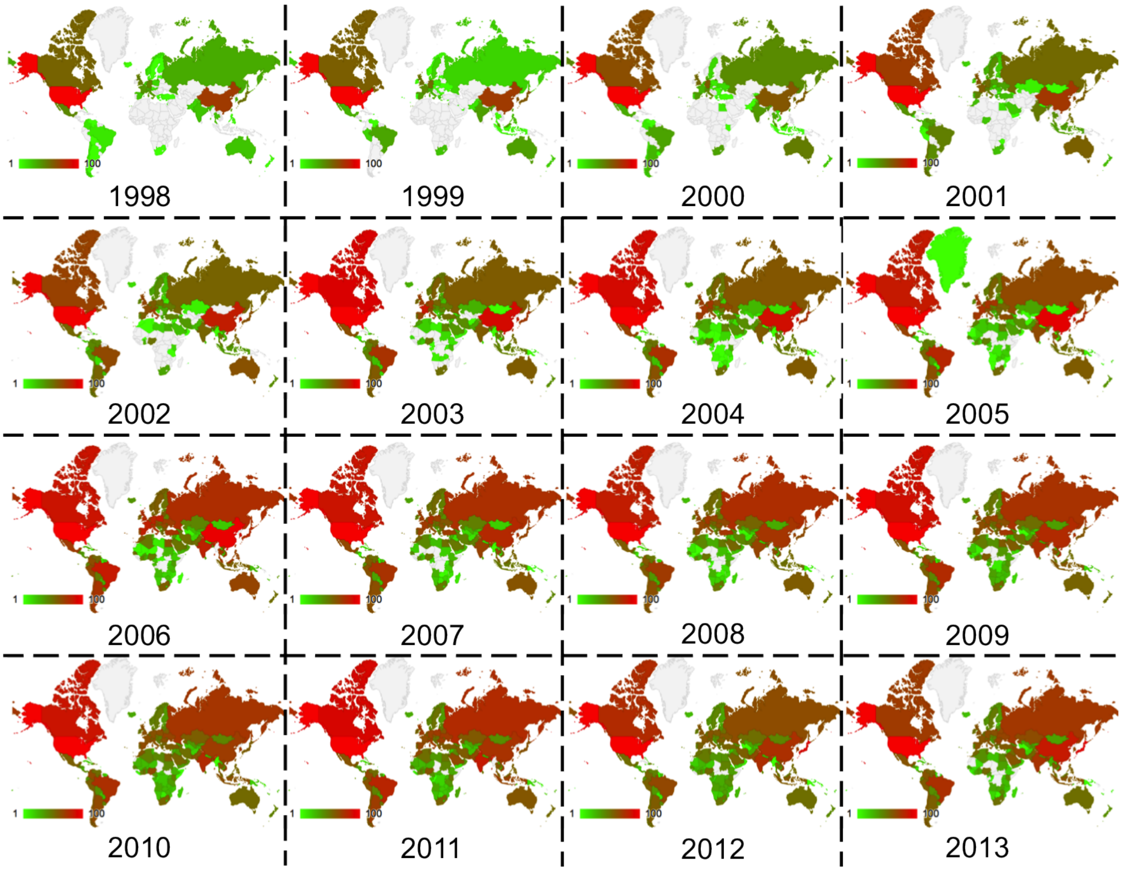

Figure 10. Geolocation distribution of senders' IP addresses (in log scale and normalized)

Fig. 9 presents the trend of average hops between sender and receiver. We observe that the number of hops was increasing over time. For instance, the average hops for 1998 was only two while it became almost eight in 2013. One possible reason is that it increased the cost for spam filters to detect or trace back the senders of email spam as spammers used more hops through intermediate proxies. It also indicates that the sender-to-receiver network becomes more complicated.

The study of header "Received" finds out the following observations (**Observation V**) :

- "RECEIVED" header is hard to be forged since it is updated along the path from the sender to the receiver.
- Spammers use more mail exchange services to avoid detection.

Fig. 10 shows the Geolocation distribution of senders' IP addresses over time. Due to space limit, we only present the Geolocation maps every two years based on the normalized number of IP addresses coming from different countries. We use the GeoIP service provided by MaxMind [55] to do the mapping between IP address and Geolocation. Also, we employ Google Geo Chart APIs [56] to implement the map drawing. The number of IP addresses from different countries has been put into log scale and then normalized into the same range from 1 to 100. Also we use green color to label countries who had the fewest sender IP address and red color to label countries who had the more sender IP addresses. White color means that no sender IP address came from the country. Observing the maps, we have the following findings in our dataset: 1) the sender IP addresses almost come from all over the world; 2) United States has the largest number of sender IP addresses along the past fifteen years; 3) Besides United States, the distribution of sender IP addresses shows dynamic changes over time. For instance, the number of sender IP addresses coming from China kept increasing until 2007 and grew again in 2013. Also, some countries had sudden increase of sender IP addresses in particular years. For example, Canada and France had sudden increase in 2003. India had sudden increase in 2011. And Japan had sudden increase in 2013. It indicates that spammers used global email service servers and also kept changing the traffic from different countries. Thus, we obtain the following observation (**Observation VI**):

- The sender IP addresses come from all over the world. United States has the largest number of

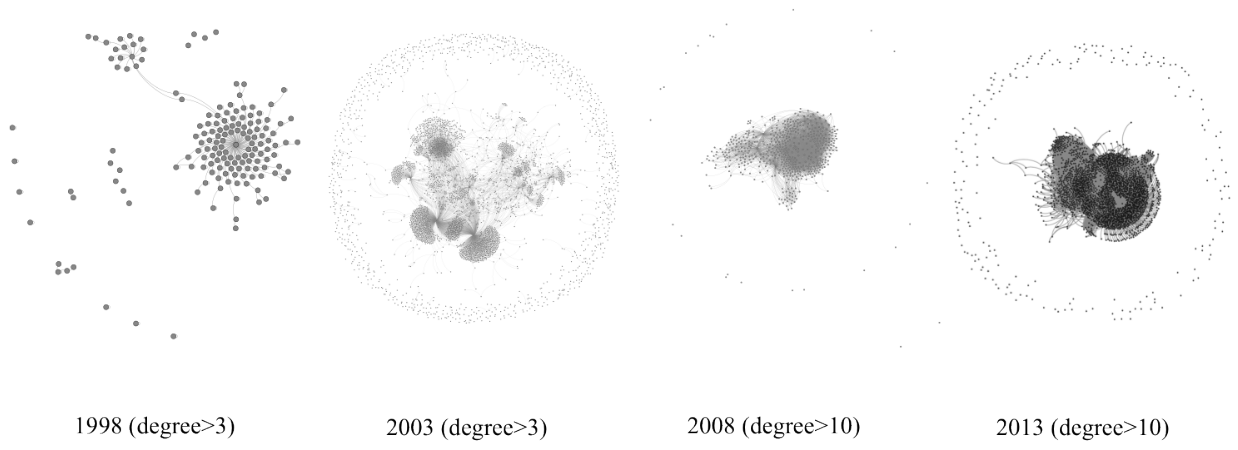

Figure 11. Sender-to-receiver routing networks every five years from 1998 to 2013

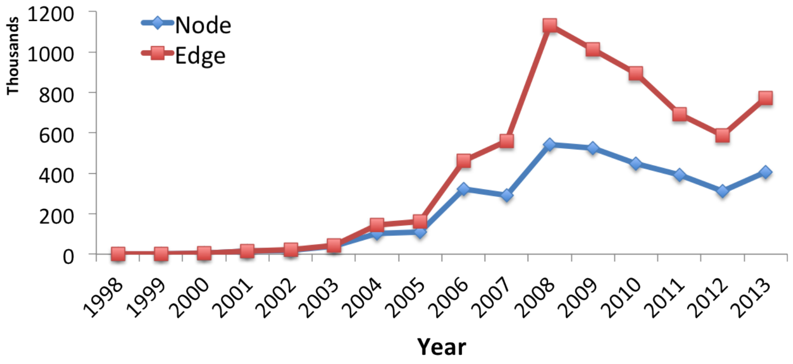

Figure 12. The trends of nodes and edges from 1998 to 2013

sender IP addressees. Generally, the distribution of IP addresses changes over time.

For the purpose of better visualization, we remove those nodes whose degree is lower than certain threshold. And also due to the space limit, we only present the network graph every five years (1998, 2003, 2008, and 2013) in Fig. 11. For 1998 and 2003, we keep the nodes whose degree is greater than 3. While for 2008 and 2013, we keep the nodes whose degree is greater than 10. The reason is that too many node overlaps occur if we choose the threshold 3 for 2008 and 2013.

Fig. 11 shows the sender-to-receiver routing network based on the IP addresses extracted from email header “Received". We observe that the complexity of graph increases explicitly along the time. For 2013, the routing network has shown much more complicated than the routing network in 2008. We also draw the trends of nodes and edges from 1998 to 2013 shown in Figure 12. In 2008, the number of nodes and edges reached the peak and later on they were decreasing to the valley in 2012. But the number of nodes and edges increased a lot in 2013. We summarize the observation as follows (**Observation VII**):

- Connection network is changing over time. Networks in recent years are more complicated than networks in earlier years.

- The number of nodes and edges have a peak value in 2008 and a valley value in 2012. The new trend is that they are both increasing.

Next, we extract the networks from our dataset for each year and use three major metrics to measure the complexity of them. The three metrics are network diameter (the longest of all the calculated shortest paths in a network), average degree (average number of edges connected to or from one node), and average clustering coefficient (a measure of degree to which nodes in a network tend to cluster together). We use those metrics to show the trend of complexity of email sender-to-receiver connection network. if the values of those metrics are large, it indicates that the complexity of

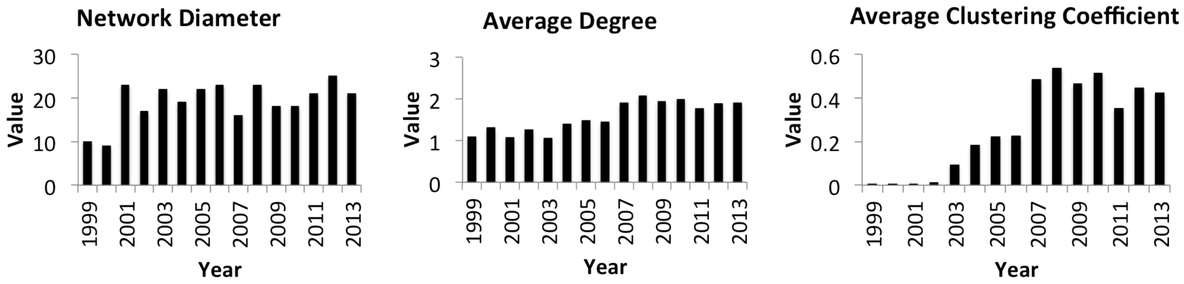

Figure 13. The comparison of three metrics from 1999 to 2013

connection network is high. The result of measurement is shown in Fig. 13.

Fig. 13 shows the three metrics comparison from 1999 to 2012. The values of them have the increasing trend overall but fluctuations existed along the time. Network diameter became more stable after 2007 and it is the same for the metrics average degree and average clustering coefficient. Those metrics kept staying at high value in terms of complexity of network. Thus, we have the following observation (**Observation VIII**):

- The complexity of networks is staying at high level and has no sign of decrease.

6. Discussion

Our large-scale evolutionary study on email spam dataset in a long period of time shows the trend of email spam business. Although the volume of email spam had a slight drop in recent years, we cannot conclude that email spam business is dying and email spam filters have won the battle against spammers. Through intensive analysis including content analysis, topic modeling and network analysis, we demonstrated that the battle is still ongoing and even worse since spammers become more sophisticated and capricious. Moreover, our study still have the following limitations and future work to do.

The dataset we used does not cover all the email spam over the fifteen years, which may influence the accuracy of our results, especially for the portion in the early years such as 1998-2000 that contains small number of email spam. Also, the bait email addresses used in data collection may cause some biases in the dataset. For example, the domain of the email address may result in that spammers forge their email addresses to the same domain.

Besides the limitation on dataset, we also have limitation on our analysis. In the topic modeling analysis, we set up the number of topics to 10 that may influence the result of topic modeling . If we change the number of topics to larger value, the result may be more accurate and fine-grained. But it should not conflict with our conclusion that the topic drift occurs frequently over time. We will take the fine-grained analysis as future work. Additionally, in the network analysis, we used the study of the header “Received" to extract sender-to-receiver network. But we cannot guarantee that no forged information exists in the header “Received". Spammers also have some techniques to spoof the header “Received" but the portion of forged headers is low since it costs spammers a lot and has certain strict requirements to meet. We will also look into the further validation work in the future.

7. Conclusions

Spam Archive dataset, which contains over 5.5 million email messages from 1998 to 2013, provides research opportunity for us to explore the real trend of email spam. In this paper, we performed a long-term (over 15 years) evolutionary study on this large scale email spam corpus. Content analysis of email spam including n-grams analysis shows the change of email content and new attacks from spammers such as legitimate URL attack and short URL camouflage. It inspires us to investigate the topic change and complexity of spamming activities. Thus, we adopted topic modeling and network analysis techniques to study topic drift and complexity of sending behaviors of spammers.

For topic modeling on email spam, we clustered the dataset based on LDA model and categorized them into ten topics: “Account Information", “Order Information", “Business News", “Sales News", “Adult Product", “Software Product", “Official News", “Free Product", “Medical Product", and “Newsletter" based on the most related terms associated. The result shows spammers changed topics over time and those topics are very attractive to users. We also found out two dominant topics, “Business News" and “Account Information", in earlier years and recent years separately. The examples we gave show that many social engineering attacks have been launched from spammers. For network analysis on complexity of spamming activities, we presented social engineering attacks from spammers by observing senders' domains. After studying the header "Received", we extracted sender IP addresses and the sender-to-receiver routing networks from the dataset. The Geolocation distribution of senders' IP addresses shows that spammers employed the servers all over the world and dynamically switched locations among different countries. Moreover, we chose three metrics: network diameter, average degree, and average clustering coefficient to measure the complexity of routing networks, showing that the sending behaviors of spammers are becoming more complicated and harder to track.

To sum up, we have obtained many new observations (Observation I-VIII). Those observations show that email spam business is becoming more sophisticated along the time and the spammers behind it evolve into more capricious in the ongoing battle with spam filters.

Acknowledgement. This research has been partially funded by National Science Foundation by CNS/SAVI (1250260), IUCRC/FRP (1127904) , CISE/CNS (1138666), RAPID (1138666), CISE/CRI (0855180), NetSE (0905493) programs, and gifts, grants, or contracts from DARPA/I2O, Singapore Government, Fujitsu Labs, and Georgia Tech Foundation through the John P. Imlay, Jr. Chair endowment. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or other funding agencies and companies mentioned above.

References

[1] S. Whittaker, V. Bellotti, and J. Gwizdka, "Email in personal information management," *ACM Communications*, vol. 49, pp. 68–73, Jan. 2006.

[2] R. Clayton, "Email traffic: a quantitative snapshot," in *the 4th Conference on Email and Anti-Spam (CEAS 2007)*, (Mountain View, CA, USA), July 2007.

[3] "DoI: Denial of Information." http://www.cc.gatech.edu/projects/doi/, 2014.

[4] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A bayesian approach to filtering junk e-mail," in *Learning for text categorization: papers from the 1998 workshop*, 1998.

[5] A. Cournane and R. Hunt, "An analysis of the tools used for the generation and prevention of spam," *Computers & Security*, vol. 23, no. 2, pp. 154 – 166, 2004.

[6] Z. Gyongyi and H. Garcia-Molina, "Web spam taxonomy," Technical Report 2004-25, Stanford InfoLab, March 2004.

[7] S. Y. Park, J.-T. Kim, and S.-G. Kang, "Analysis of applicability of traditional spam regulations to voip spam," in *Advanced Communication Technology, 2006. ICACT 2006. The 8th International Conference*, vol. 2, pp. 3 pp.–1217, 2006.

[8] P. Hayati, V. Potdar, A. Talevski, N. Firoozeh, S. Sarenche, and E. Yeganeh, "Definition of spam 2.0: New spamming boom," in *Proceedings of the 4th IEEE International Conference on Digital Ecosystems and Technologies (DEST)*, pp. 580–584, 2010.

[9] L. F. Cranor and B. A. LaMacchia, "Spam!," *ACM Communications*, vol. 41, pp. 74–83, Aug. 1998.

[10] MAAWG, "Email Metrics Report 2011," tech. rep., November 2011.

[11] J. Goodman, G. V. Cormack, and D. Heckerman, "Spam and the ongoing battle for the inbox," *ACM Communications*, vol. 50, pp. 24–33, Feb. 2007.

[12] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma, "Adversarial classification," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, (New York, NY, USA), pp. 99–108, ACM, 2004.

[13] D. Chinavle, P. Kolari, T. Oates, and T. Finin, "Ensembles in adversarial classification for spam," in *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, (New York, NY, USA), pp. 2015–2018, ACM, 2009.

[14] B. Biggio, G. Fumera, and F. Roli, "Evade hard multiple classifier systems," in *Applications of Supervised and Unsupervised Ensemble Methods* (O. Okun and G. Valentini, eds.), vol. 245 of *Studies in Computational Intelligence*, pp. 15–38, Springer Berlin Heidelberg, 2009.

[15] C. Pu and S. Webb, "Observed trends in spam construction techniques: A case study of spam evolution," in *Proceedings of the Third Conference on Email and Anti-Spam (CEAS 2006)*, (Mountain View, CA, USA), July 2006.

[16] T. Fawcett, ""in vivo" spam filtering: a challenge problem for kdd," *SIGKDD Explor. Newsl.*, vol. 5, pp. 140–148, Dec. 2003.

[17] D. Irani, S. Webb, J. Giffin, and C. Pu, "Evolutionary study of phishing," *eCrime Researchers Summit, 2008*, pp. 1–10, 2008.

[18] D. Wang, D. Irani, and C. Pu, "Evolutionary study of web spam: Webb spam corpus 2011 versus webb spam corpus 2006," in *Proceedings of the 8th International Conference on Collaborative Computing: Networking, Applications and Work-sharing (CollaborateCom)*, (Pittsburgh, PA, USA), pp. 40–49, October 2012.

[19] P. Guerra and D. Guedes, "Exploring the spam arms race to characterize spam evolution," in *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS 2010)*, (Redmond, Washington USA), July 2010.

[20] "Untroubled dataset website." http://untroubled.org/spam/, 2014.

[21] "Kaspersky lab." http://usa.kaspersky.com/, 2013.

[22] K. Rapoza, "The dying business of email spam." http://usa.kaspersky.com/about-us/press-center/in-the-news/dying-business-email-spam, 2012.

[23] D. Gudkova, "Kaspersky security bulletin: Spam evolution 2012." http://www.securelist.com/en/analysis/204792276/Kaspersky_Security_Bulletin_Spam_Evolution_2012, 2012.

[24] X. Carreras and L. Marquez, "Boosting trees for anti-spam email filtering," *arXiv preprint cs/0109015*, 2001.

[25] R. Wang, A. Youssef, and A. Elhakeem, "On Improving the Performance of Spam Filters Using Heuristic Feature Selection Techniques," in *Proceedings of 23rd Biennial Symposium on Communications, 2006*, pp. 227–230, Ieee, 2006.

[26] J. Chan, I. Koprinska, and J. Poon, “Co-training with a Single Natural Feature Set Applied to Email Classification,” in *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence (WI’04)*, pp. 586–589, Ieee, 2004.

[27] W. Liu and T. Wang, “Multi-field learning for email spam filtering,” in *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, (New York, NY, USA), p. 745, ACM Press, 2010.

[28] D. Sculley and G. Wachman, “Relaxed online SVMs for spam filtering,” in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 415–422, April 2007.

[29] J. Attenberg, K. Weinberger, and A. Dasgupta, “Collaborative Email-Spam Filtering with the Hashing Trick,” in *CEAS*, pp. 1–4, 2009.

[30] M. Balakumar and V. Vaidehi, “Ontology based classification and categorization of email,” in *Proceedings of Signal Processing, Communications and Networking*, pp. 199–202, 2008.

[31] A. Dasgupta, M. Gurevich, and K. Punera, “Enhanced email spam filtering through combining similarity graphs,” in *Proceedings of the fourth ACM international conference on Web search and data mining*, (New York, NY, USA), p. 785, ACM Press, 2011.

[32] J. Jung and E. Sit, “An empirical study of spam traffic and the use of DNS black lists,” in *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, (New York, NY, USA), p. 370, ACM Press, 2004.

[33] A. Ramachandran, N. Feamster, and S. Vempala, “Filtering spam with behavioral blacklisting,” in *Proceedings of the 14th ACM conference on Computer and communications security*, (New York, NY, USA), p. 342, ACM Press, 2007.

[34] M. Xie, H. Yin, and H. Wang, “An effective defense against email spam laundering,” in *Proceedings of the 13th ACM conference on Computer and communications security*, (New York, NY, USA), p. 179, ACM Press, 2006.

[35] L. Zhuang, J. Dunagan, D. Simon, and H. Wang, “Characterizing Botnets from Email Spam Records.,” in *Proceedings of the first USENIX workshop on large-scale exploits and emergent threats (LEET 08)*, 2008.

[36] S. Webb, J. Caverlee, and C. Pu, “Introducing the webb spam corpus: Using email spam to identify web spam automatically,” in *Proceedings of the Third Conference on Email and Anti-Spam (CEAS 2006)*, (Mountain View, CA, USA), July 2006.

[37] D. Wang, D. Irani, and C. Pu, “A social-spam detection framework,” in *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS 2011)*, (Perth, Australia), pp. 46–54, September 2011.

[38] C. Bird, A. Gourley, and P. Devanbu, “Mining email social networks,” in *the 2006 international workshop on Mining software repositories*, pp. 137–143, 2006.

[39] A. McCallum, X. Wang, and A. Corrada-Emmanuel, “Topic and role discovery in social networks with experiments on enron and academic email.,” *J. Artif. Intell. Res.(JAIR)*, vol. 30, pp. 249–272, 2007.

[40] A. Culotta, R. Bekkerman, and A. McCallum, “Extracting social networks and contact information from email and the web,” in *Proceedings of the First Conference on Email and Anti-Spam (CEAS 2004)*, 2004.

[41] D. Wang, D. Irani, and C. Pu, “A study on evolution of email spam over fifteen years,” in *Collaborative Computing: Networking, Applications and Worksharing (Collaboratecom), 2013 9th International Conference Conference on*, pp. 1–10, Oct 2013.

[42] D. Wang, “Analysis and detection of low quality information in social networks,” in *Proceedings of Ph.D. Symposium at 30th IEEE International Conference on Data Engineering (ICDE 2014)*, (Chicago, IL, United States), 2014.

[43] D. Wang, D. Irani, and C. Pu, “A perspective of evolution after five years: A large-scale study of web spam evolution,” *Int. J. Cooperative Inf. Syst.*, vol. 23, no. 2, 2014.

[44] Y. Chung, *A Study on the Evolution and Emergence of Web Spam*. PhD thesis, Univ. of Tokyo, Tokyo, Japan, 2011.

[45] D. Fetterly, M. Manasse, M. Najork, and J. Wiener, “A large-scale study of the evolution of web pages,” in *Proceedings of the 12th international conference on World Wide Web*, WWW ’03, (New York, NY, USA), pp. 669–678, 2003.

[46] “Electronic communications and transactions act, 2002.” `http://www.internet.org.za/ect_act.html`, 2002.

[47] “Multipurpose internet mail extensions (MIME) part one: Format of internet message bodies.” `http://tools.ietf.org/html/rfc2045`, 1996.

[48] “Text::Ngrams - flexible ngram analysis (for characters, words, and more).” `http://search.cpan.org/dist/Text-Ngrams/Ngrams.pm`, 2014.

[49] “Defensive aids sub system (DASS).” `http://www.eurofighter.com/capabilities/technology/sensor-fusion/defensive-aids-sub-system.html`, 2014.

[50] A. K. McCallum, “MALLET: A machine learning for language toolkit.” `http://mallet.cs.umass.edu`, 2002.

[51] D. M. Blei, A. Ng, and M. Jordan, “Latent dirichlet allocation,” *JMLR*, vol. 3, pp. 993–1022, 2003.

[52] S. Hao, N. A. Syed, N. Feamster, A. G. Gray, and S. Krasser, “Detecting spammers with snare: Spatio-temporal network-level automatic reputation engine,” in *Proceedings of the 18th Conference on USENIX Security Symposium*, SSYM’09, pp. 101–118, 2009.

[53] “Python: email – an email and MIME handling package.” `http://docs.python.org/2/library/email`, 2014.

[54] “Gephi: an open source graph visualization and manipulation software.” `http://gephi.org`, 2014.

[55] “MaxMind – IP geolocation and online fraud prevention.” `http://www.maxmind.com/en/home`, 2014.

[56] “Visualization: Geochart – Google charts – Google developers.” `https://developers.google.com/chart/interactive/docs/gallery/geochart`, 2014.